REMOVAL SAMPLING: A GENERALIZATION WITH SUBPOPULATIONS AND SUBSAMPLING

BU-635-M

David B. Hall

December, 1977

Abstract

In a closed population with unknown size, a sample is taken with each element having equal unknown probability p of being in the sample. With p constant over samples, repeated samples without replacement allow the estimation of p and population size N . If each sample is subsampled with the probability of an element being in the subsample being θ, a known constant, and the subsample then divided into distinct subpopulations, then the following formulas are used to estimate subpopulation sizes N_i and p by maximum likelihood techniques:

$$\sum \hat{N}_i = \theta^{-1} \frac{s_{..}}{1-(1-\hat{p})^m}$$

$$\hat{p}^{-1}(1-\hat{p}) - \left[1-(1-\hat{p})^m\right]^{-1} m(1-\hat{p})^m = s_{..}^{-1} \sum_{i=1}^{m} (i-1)s_{i.}$$

$$\sum \hat{N}_i s_{.j} = \hat{N}_j s_{..}$$

$(s_{ij}$ = number of elements from subpopulation j in subsample i) .

An asymptotic variance-covariance matrix, V^*, for the $\hat{N}_j$'s is derived. With average individual biomass within subpopulation j estimated by $\hat{w}_j$ and D_w as the variance-covariance matrix of these estimates, $\sum_j \hat{w}_j \hat{N}_j$ estimates total biomass, with sampling variance: $\underline{N}'D_w\underline{N} + \underline{w}'V^*\underline{w} + tr(V^*D_w)$. $[\underline{N} = (N_1, N_2, \cdots, N_m), \underline{w} = (w_1, w_2, \cdots, w_m)]$.

This problem was examined for $\theta = 1$ and $m = 1$ by Moran (1951) and Zippin (1956, 1958).

REMOVAL SAMPLING: A GENERALIZATION WITH SUBPOPULATIONS AND SUBSAMPLING

David B. Hall

1. Introduction

The removal method of sampling and maximum likelihood estimation of population number using the data obtained by removal sampling are described in Moran (1951) and Zippin (1956, 1958). In a closed animal population a series of m catches provides the data. On each catch the animals have an equal, independent chance of being caught. When an animal has been caught it is removed from the population. The assumptions which must be met before using the removal method are discussed in Seber (1973, pp. 311-315):

"(1) The population is closed.

(2) The probability of capture in the ith sample is the same for each individual exposed to capture.

(3) The probability of capture p remains constant from sample to sample."

Let p represent the probability of being caught, c_i represent the number of animals in catch i, and N represent the population size. The conditional probability of catching c_j animals in the jth catch, given $c_1, c_2, \cdots, c_{j-1}$, would be

$$\binom{N-c_1-c_2\cdots c_{j-1}}{c_j} p^{c_j}(1-p)^{N-\sum_{i=1}^{m} c_i} .$$

The likelihood equation is the product of these conditional probabilities as j varies from 1 to m . The resulting estimates of p and N are the solution to the equations:

$$\left(\sum_{j=1}^{m} c_j\right)\hat{p}^{-1} - \left[m\hat{N} - \sum_{j=1}^{m}(m-j+1)c_j\right](1-\hat{p})^{-1} = 0 \qquad (1)$$

and

$$m \log(1-\hat{p}) - \log\left[\left(\hat{N} - \sum_{j=1}^{m} c_j\right)\Big/\hat{N}\right] = 0 \quad . \tag{2}$$

A researcher raising Daphnia magna, a planktonic animal, in closed laboratory populations wished to estimate total biomass of a population using a number of size categories. With estimates of number of animals and average weight of animals in each size category total biomass was estimated. The number of animals in each size category was estimated using removal sampling with subsampling. For each catch a subsample was taken with known probability θ of a caught animal being in the subsample. The estimated average weight and its estimated variance for each size category were calculated after random sampling from the general population of animals in the species of interest.

Two practical problems lead to the use of a more general form for the removal method. If the catches are too large, it may be necessary to subsample each catch, still removing the entire catch from the population. The data in some size category, due to small numbers in the population, may provide highly unreliable estimates. This problem may be handled when the assumption of equal probability of capture for each individual in the entire population is acceptable. This assumption allows estimation of each subpopulation using only marginal values: the total number counted in each catch and the total number counted from each subpopulation.

2. Estimation of Population Size

Notation: c_{ij} = number of individuals from subpopulation j in catch i $(i = 1, 2, \cdots, m)(j = 1, 2, \cdots, k)$.

s_{ij} = number of individuals from subpopulation j in the counted subsample from catch i .

$s_{i\cdot}$ = total number of individuals in the subsample from catch i .

$s_{\cdot j}$ = total number of individuals counted from subpopulation j .

p = probability of capture in catch i for all individuals not already caught.

θ = probability of being counted for each individual which has been caught (known).

N_j = number of individuals in subpopulation j .

$N_{\cdot}$ = total population.

$s_{\cdot\cdot}$ = total number of individuals counted = $\sum_{i=1}^{m} \sum_{j=1}^{k} s_{ij}$.

The likelihood equation with the observed variables (s_{ij}'s) and the unobserved variables (c_{ij}'s) can be written as the product of likelihood equations for each subpopulation. $L' = L_1' \times L_2' \times \cdots \times L_k'$.

$$L_j' = \binom{N_j}{c_{1j}} p^{c_{1j}} (1-p)^{N_j - c_{1j}} \binom{N_j - c_{1j}}{c_{2j}} p^{c_{2j}} (1-p)^{N_j - c_{1j} - c_{2j}} \cdots \binom{N_j - \sum_{i=1}^{m-1} c_{ij}}{c_{mj}} p^{c_{mj}} (1-p)^{N_j - c_{\cdot j}}$$

$$\times \binom{c_{1j}}{s_{1j}} \theta^{s_{1j}} (1-\theta)^{c_{1j}-s_{1j}} \binom{c_{2j}}{s_{2j}} \theta^{s_{2j}} (1-\theta)^{c_{2j}-s_{2j}} \cdots \binom{c_{mj}}{s_{mj}} \theta^{s_{mj}} (1-\theta)^{c_{mj}-s_{mj}}$$

$$= \frac{N_j!}{s_{1j}!s_{2j}!\cdots s_{mj}!(c_{1j}-s_{1j})!\cdots(c_{mj}-s_{mj})!\left(N_j - \sum_{i=1}^{m} c_{ij}\right)!} p^{\sum_{i=1}^{m} c_{ij}} (1-p)^{mN_j - \sum_{i=1}^{m} (m-i+1)c_{ij}}$$

$$\times \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{m} s_{ij}} \left(1-\theta\right)^{\sum_{i=1}^{m} c_{ij}} \quad .$$

Summing over all possible values for the c_{ij}'s to eliminate the unobservable variables results in the equation:

$$L_j = \frac{N_j!}{s_{1j}!\cdots s_{mj}!(N_j - s_{\cdot j})!} p^{s_{\cdot j}} (1-p)^{\sum_{i=1}^{m} (i-1)s_{ij}} [1-\theta+\theta(1-p)^m]^{N_j - s_{\cdot j}} \quad .$$

The likelihood equation using only the observed variables is the product of the L_j's .

$$L = \frac{N_1!N_2!\cdots N_k!}{\left(\prod_{i,j} s_{ij}!\right)\left(\prod_j (N_j - s_{\cdot j})!\right)} p^{s_{\cdot\cdot}} (1-p)^{\sum_{i=1}^{m}(i-1)s_{i\cdot}} \left(1-\theta+\theta(1-p)^m\right)^{N_\cdot - s_{\cdot\cdot}} .$$

This may be factored into the conditional likelihood of the $s_{\cdot j}$'s given $s_{\cdot\cdot}$, the likelihood of the $s_{i\cdot}$'s, and a parameterless function of the s_{ij}'s .

$$L = \left[\frac{\binom{N_1}{s_{\cdot 1}}\binom{N_2}{s_{\cdot 2}}\cdots\binom{N_k}{s_{\cdot k}}}{\binom{N_\cdot}{s_\cdot}}\right]$$

$$\times \left[\frac{N_\cdot!}{s_{1\cdot}!s_{2\cdot}!\cdots s_{m\cdot}!\left(N_\cdot - s_{\cdot\cdot}\right)!} p^{s_{\cdot\cdot}} (1-p)^{\sum_{i=1}^{m}(i-1)s_{i\cdot}} \left(1-\theta+\theta(1-p)^m\right)^{N_\cdot - s_{\cdot\cdot}}\right]$$

$$\times \left[\frac{\prod_j (s_{\cdot j})!\prod_i (s_{i\cdot})!}{s_{\cdot\cdot}!\prod_i\prod_j s_{ij}!}\right] .$$

This factorization illustrates the fact that the ML estimates of the parameters will be functions of the marginal values only. Also, the $\hat{N}_j$'s will be in the same relationship to each other as the $s_{\cdot j}$'s . That is, $\hat{N}_j s_{\cdot j'} = \hat{N}_{j'} s_{\cdot j}$ for all j and j' .

While the partial derivative with respect to p presents no problem, the partial with respect to N_j is usually approximated because of the $(N_j + c)!$ terms in the likelihood equation. Two methods of approximation which achieve the same result are:

(i) Estimate $(N_j+c)!$ using Stirling's formula $\left(n! = (2\pi)^{\frac{1}{2}} e^{-n} n^{n+\frac{1}{2}}\right)$ and differentiate, resulting in

$$\frac{\partial \ln L}{\partial N_j} \doteq \ln N_j - \ln(N_j - s_{\cdot j}) + \tfrac{1}{2}\left(N_j^{-1} - (N_j - s_{\cdot j})^{-1}\right) + \ln\left(1-\theta+\theta(1-p)^m\right) \quad .$$

In practice the term $\frac{1}{2}\left(N_j^{-1} - (N_j - s_{\cdot j})^{-1}\right)$ is ignored.

(ii) Estimate the derivative with the first order difference equation $(h=1)$,

$$\frac{\partial \ln L}{\partial N_j} \doteq \Delta \ln L(N_j) = \ln N_j - \ln(N_j - s_{\cdot j}) + \ln\left(1-\theta+\theta(1-p)^m\right) \quad .$$

In practice these two approaches yield identical approximations to $\partial \ln L/\partial N_j$.

Using the fact that $N_j!/[(N_j - s_{\cdot j})!] = N_j(N_j-1)\cdots(N_j - s_{\cdot j}+1)$, the log likelihood equation can be written as

$$\ln L = \sum_{j=1}^{k}\left[\sum_{c=0}^{s_{\cdot j}-1} \ln(N_j - c) + \sum_{i=1}^{m} \ln\left(\frac{1}{s_{i\cdot}!}\right)\right] + s_{\cdot\cdot}\ln p + \sum_{i=1}^{m}(i-1)s_{i\cdot}\ln(1-p)$$
$$+ (N_{\cdot} - s_{\cdot\cdot})\ln\left(1-\theta+\theta(1-p)^m\right) \quad .$$

The first derivative with respect to N_j is:

$$\frac{\partial \ln L}{\partial N_j} = \sum_{c=0}^{s_{\cdot j}-1} \frac{1}{N_j - c} + \ln\left(1-\theta+\theta(1-p)^m\right) \quad (j=1, 2, \cdots, k) \quad .$$

Setting this and $\partial \ln L/\partial p$ equal to zero results in the following equations for estimation:

$$1 - e^{\sum_{c=0}^{s_{\cdot\cdot}-1}(\hat{N}_{\cdot} - c)^{-1}} = \theta\left(1-(1-p)^m\right) \quad . \tag{3E}$$

$$\left(1-\theta+\theta(1-\hat{p})^m\right)\left(s_{..}(1-\hat{p}) - \sum_{i=1}^{m}(i-1)s_{i.}\hat{p}\right) = \hat{p}(1-\hat{p})(\hat{N}_{.}-s_{..})\left(m\theta(1-\hat{p})^{m-1}\right) . \quad (4E)$$

$$\hat{N}_{.}s_{.j} = \hat{N}_j s_{..} \quad (j=1, 2, \cdots, k) . \quad (5E)$$

Using the formula

$$\sum_{i=a}^{b}\frac{1}{i} \doteq \int_{a-.418}^{b+.582}\frac{1}{x}\,dx \cdot \left[\int_{1-.418}^{1.582}\frac{1}{x}\,dx = 1\right]$$

$$= \ln(b+.582) - \ln(a-.418)$$

equation (3E) becomes

$$\ln(\hat{N}_{..}+.582) - \ln(\hat{N}_{..}-s_{..}+.582) + \ln\left(1-\theta+\theta(1-\hat{p})^m\right) = 0$$

which simplifies to

$$\hat{N}_{..} = \frac{s_{..}}{\theta\left(1-(1-\hat{p})^m\right)} + 0.582 . \quad (3E.1)$$

Substituting for $\hat{N}_{..}$ in (4E) results in the equation

$$s_{..}\left((1-\hat{p}) - \frac{m\hat{p}(1-\hat{p})^m}{1-(1-\hat{p})^m}\right) - \sum_{i=1}^{m-1} i\; s_{(i+1).}\hat{p} = \frac{(0.582)m\theta\hat{p}(1-\hat{p})^m}{1-\theta\left(1-(1-\hat{p})^m\right)} . \quad (4E.1)$$

Using first order difference equations for the N_j's and the first partial derivative for p on the log-likelihood results in the following equations:

$$\Delta\ln L(N_j) = \ln(N_j) - \ln(N_j-s_{.j}) + \ln\left(1-\theta+\theta(1-\hat{p})^m\right) \quad (j=1, 2, \cdots, k)$$

and

$$\frac{\partial\ln L}{\partial p} = s_{..}p^{-1} - (1-p)^{-1}\sum_{i=1}^{m}(i-1)s_{i.} + \left(1+\theta+\theta(1-p)^m\right)^{-1}(N_{.}-s_{..})\left(-m\theta(1-\hat{p})^{m-1}\right) .$$

Setting these equal to zero results in the following equations for estimation:

$$\hat{N}_{.} = \theta^{-1} \frac{s_{..}}{1-(1-\hat{p})^{m}} , \tag{3}$$

$$\hat{p}^{-1}(1-\hat{p}) - \left(1-(1-\hat{p})^{m}\right)^{-1} m(1-\hat{p})^{m} = s_{..}^{-1} \sum_{i=1}^{m} (i-1)s_{i.} , \tag{4}$$

and

$$\hat{N}_{.} s_{.j} = \hat{N}_{j} s_{..} \quad . \qquad (j = 1, 2, \cdots, k) \tag{5}$$

Note that (3) and (3E.1) differ only by the correction factor, 0.582, and therefore differ by at most one when rounded to the nearest integer, when $\hat{p}$ is constant. Note that the only difference between (4) and (4E.1) is the correction factor, $\{[(0.582)m\theta\hat{p}(1-\hat{p})^{m}]/[1-\theta(1-(1-\hat{p})^{m})]\}$. Estimates may be found by solving (4E.1) and substituting into (3E.1) or by solving (4) and substituting into (3). The solutions should not differ by more than one in the estimation of $\hat{N}_{..}$.

Two special cases are to be considered. When $\theta = 1$ and $k = 1$ the estimation equations (3), (4), and (5) reduce to

$$\ln \frac{\hat{N}}{\hat{N}-s_{.}} + \ln\left[(1-p)^{m}\right] = 0 , \tag{6}$$

$$s_{.}\hat{p}^{-1} - (1-\hat{p})^{-1} \sum_{i=1}^{m} (i-1)s_{i} + (1-\hat{p})^{-m}(\hat{N}-s_{.})\left(-m(1-\hat{p})^{m-1}\right) = 0 ,$$

and

$$s_{.}\hat{p}^{-1} - (1-\hat{p})^{-1}\left[\sum_{i=1}^{m} (i-1)s_{i} + m(\hat{N}-s_{.})\right] = 0 \quad . \tag{7}$$

Equations (6) and (7) can be seen to be equivalent to equations (1) and (2). $m = 3$ is a popular choice for number of catches because of the closed solution for estimation of parameters and because additional catches are usually not cost-effective in improving precision. When $m = 3$ the estimators are:

$$\hat{N}_{\cdot} = \theta^{-1} \frac{6X^2-3XY-Y^2+Y(Y^2+6XY-3X^2)^{\frac{1}{2}}}{18(X-Y)} \tag{8}$$

$$\hat{p} = \frac{3X-Y-(Y^2+6XY-3X^2)^{\frac{1}{2}}}{2X} \tag{9}$$

$$\hat{N}_j = \frac{s_{\cdot j}}{s_{\cdot\cdot}} \hat{N} \quad . \tag{10}$$

where $X = 2s_{1\cdot} + s_{2\cdot}$ and $Y = s_{1\cdot} + s_{2\cdot} + s_{3\cdot}$.

3. Sampling Variance Estimation for Population Size

The variance-covariance matrix for maximum likelihood estimates is asymptotically

$$V = \left[\left\{-E \frac{\partial^2 \ln L}{\partial\beta_i \partial\beta_j}\right\}\right]^{-1} \quad .$$

In this case the elements of V^{-1} are:

$$-E \frac{\partial \Delta \ln L(N_j)}{\partial N_j} = -E\left[\frac{1}{N_j} - \frac{1}{N_j - s_{\cdot j}}\right] ,$$

$$-E \frac{\partial \Delta \ln L(N_j)}{\partial N_{j'}} = 0 \text{ for all } j \neq j' ,$$

$$-E \frac{\partial^2 \ln L}{\partial p \partial N_j} = -E \frac{\partial \Delta \ln L(N_j)}{\partial p} = -E\left[\frac{-m\theta(1-p)^{m-1}}{1-\theta+\theta(1-p)^m}\right] ,$$

$$-E \frac{\partial^2 \ln L}{\partial p^2}$$

$$= -E\left[-p^{-2}s_{\cdot\cdot} - (1-p)^{-2}\sum_{i=1}^{m}(i-1)s_{i\cdot} + (N_{\cdot} - s_{\cdot\cdot})\left(\frac{m(m-1)\theta(1-p)^{m-2}}{1-\theta+\theta(1-p)^m} - \frac{m^2\theta^2(1-p)^{2(m-1)}}{\left(1-\theta+\theta(1-p)^m\right)^2}\right)\right] .$$

Using the expected values $E(s_{i.}) = N_{.}p(1-p)^{i-1}\theta$, $E(s_{.j}) = N_j\theta\left(1-(1-p)^m\right)$, and $E(s_{..}) = N_{.}\theta\left(1-(1-p)^m\right)$ and expanding $1/(N_j-s_{.j})$ around $1/(N_j-Es_{.j})$ results in:

$$a = -E\,\frac{\partial^2 \ln L}{\partial p^2}$$

$$= N_{.}\theta\left[p^{-2}\left(1-(1-p)^m\right)+\sum_{i=1}^{m}(i-1)p(1-p)^{i-3}-m(m-1)(1-p)^{m-2}+\theta m^2(1-p)^{2m-2}\left(1-\theta+\theta(1-p)^m\right)^{-1}\right],$$

$$b = -E\,\frac{\partial^2 \ln L}{\partial N_j \partial p} = m\theta(1-p)^{m-1}\left(1-\theta+\theta(1-p)^m\right)^{-1},$$

$$d_i = -E\,\frac{\partial \Delta \ln L(N_i)}{\partial N_i} = \theta\left(1-(1-p)^m\right)\left[\left[N_j\left(1-\theta+\theta(1-p)^m\right)\right]^{-1}-\left[N_j\left(1-\theta+\theta(1-p)^m\right)\right]^{-2}\right].$$

The exact value for $\partial^2 \ln L/\partial N_j^2$ is

$$-\sum_{c=0}^{s_{.j}-1}\left(N_j-c\right)^{-2}.$$

Thus, the exact value for d_i is

$$d_i^e = -E\,\frac{\partial^2 \ln L}{\partial N_j^2} = +E\sum_{c=0}^{s_{.j}-1}\left(N_j-c\right)^{-2}$$

$$= \sum_{c=0}^{N_j-1}\frac{1}{(N_j-c)^2}\Pr(s_{.j} \geq c+1),$$

$$\Pr(s_{.j} \geq c+1) = \sum_{b=c+1}^{N_j}\sum_{s_{1j}}\frac{N_j!}{s_{1j}!s_{2j}!\cdots s_{mj}!(N_j-b)!}p^b(1-p)^{\sum_{i=1}^{m}(i-1)s_{ij}}\left(1-\theta+\theta)1-p)^m\right)^{N_j-b}$$

$$= \sum_{b=c+1}^{N_j}\binom{N_j}{b}p^b\left(1-\theta+\theta(1-p)^m\right)^{N_j-b}\sum_{s_{ij}}\frac{b!}{s_{1j}!s_{2j}!\cdots s_{mj}!}(1-p)^{\sum_{i=1}^{m}(i-1)s_{ij}}.$$

By the multinomial theorem

$$\sum_{s_{ij}} \frac{b!}{s_{1j}!s_{2j}!\cdots s_{mj}!}(1-p)^{\sum_{i=1}^{m}(i-1)s_{ij}} = \left(\sum_{i=0}^{m-1}(1-p)^i\right)^b ,$$

so

$$\Pr(s_{\cdot j} \geq c+1) = \sum_{b=c+1}^{N_j}\binom{N_j}{b}p^b\left(1-\theta+\theta(1-p)^m\right)^{N_j-b}\left(\sum_{a=0}^{m-1}(1-p)^a\right)^b .$$

Thus

$$d_i^e = \sum_{c=0}^{N_j-1}\sum_{b=c+1}^{N_j}\left(\sum_{a=0}^{m-1}(1-p)^a\right)^b\left(\frac{1}{(N_j-c)^2}\right)\binom{N_j}{b}p^b\left(1-\theta+\theta(1-p)^m\right)^{N_j-b} .$$

Using the facts that

$$\left(\sum_{i=0}^{r-1}x^i = \frac{1-x^r}{1-x}\right)$$

and

$$\left(\sum_{i=1}^{n} i^{-2} \doteq \int_{\frac{1}{2}}^{n+\frac{1}{2}} x^{-2}dx \doteq 2 \text{ for large } n\right) ,$$

d_i^e reduces to

$$d_i^e \doteq \sum_{x=1}^{N_j} N_j x^{-2}\left[\left(1-(1-p)^m\right)^{-1}-\theta\right]^x\left(1-(1-p)^m\right)^{N_j+1}\left[(1-p)^m-\theta\left(1-(1-p)^m\right)\right]^{-1}$$
$$- 2N_j\left(1-(1-p)^m\right)^{N_j+1}\left[(1-p)^m-\theta\left(1-(1-p)^m\right)\right]^{-1} .$$

$$= N_j\left(1-(1-p)^m\right)^{N_j+1}\left[(1-p)^m-\theta\left(1-(1-p)^m\right)\right]^{-1}\left[\sum_{x=1}^{N_j}x^{-2}\left[\left(1-(1-p)^m\right)^{-1}-\theta\right]^x-2\right]$$

$$= N_j\left(1-(1-p)^m\right)^{N_j+1}\left[(1-p)^m-\theta\left(1-(1-p)^m\right)\right]^{-1}\sum_{x=1}^{N_j}x^{-2}\left[\left[\left(1-(1-p)^m\right)^{-1}-\theta\right]^x-1\right] .$$

Since the properties of the variance estimate are asymptotic, the approximation introduced by the use of d_i rather than d_i^e should not be important.

Defining D as a k x k matrix such that $d_{ij} = d_i$ or d_i^e if i = j and zero elsewhere allows V^{-1} to be written as the partitioned matrix

$$\begin{bmatrix} a & b\mathbf{1}' \\ b\mathbf{1} & D \end{bmatrix} .$$

Using the formulas for inversion of a partitioned matrix (Searle, p. 210), V is:

$$\begin{bmatrix} (a - b^2\mathbf{1}'D^{-1}\mathbf{1})^{-1} & -ab\mathbf{1}'D^{-1} \\ -abD^{-1}\mathbf{1} & D^{-1} + b^2(a - b^2\mathbf{1}'D^{-1}\mathbf{1})^{-1}D^{-1}\mathbf{1}\mathbf{1}'D^{-1} \end{bmatrix}$$

where $\mathbf{1}$ is a vector of k ones.

The k x k variance-covariance matrix for the $\hat{N}_j$'s will be designated V^* with elements v_{ij} .

$$v_{ij} = \begin{cases} d_i^{-1} + d_i^{-2}b^2\left(a - b^2 \sum_{i=1}^{k} d_i^{-1}\right)^{-1} & \text{if} \quad i = j, \\ d_i^{-1}d_j^{-1}b^2\left(a - b^2 \sum_{i=1}^{k} d_i^{-1}\right)^{-1} & \text{if} \quad i \neq j \quad . \end{cases}$$

To estimate the v_{ij}'s the N_j's and p in a, b, and d_i's are replaced by their estimates.

As a check consider V when $\theta = 1$ and $k = 1$. The terms become:

$$a = N_{\bullet}\left[p^{-2}\left(1-(1-p)^m\right)+\sum_{i=1}^{m}(i-1)p(1-p)^{i-3}-m(m-1)(1-p)^{m-2}+m^2(1-p)^{2m-2}(1-p)^{-m}\right]$$

$$= N_{\bullet}\left[p^{-2}\left(1-(1-p)^m\right)+\sum_{i=1}^{m}(i-1)p(1-p)^{i-3}+m(1-p)^{m-2}\right]$$

$$= N_{\bullet}\left[p^{-2}\left(1-(1-p)^m\right)+\sum_{i=1}^{m}\left[(i-1)p(1-p)^{i-3}+(1-p)^{m-2}\right]\right],$$

$$b = m(1-p)^{m-1}(1-p)^{-m} = \frac{m}{1-p}$$

$$d_i = d = \left(1-(1-p)^m\right)\left[\left(N(1-p)^m\right)^{-1}-\left(N(1-p)^m\right)^{-2}\right]$$

and

$$-\frac{\partial\Delta\ln L(N)}{\partial N} = \frac{1}{N-s_{\bullet}} - \frac{1}{N}$$

$$-\frac{\partial^2\ln L}{\partial p\partial N} = m(1-p)^{m-1}(1-p)^{-m} = \frac{m}{1-p}$$

$$-\frac{\partial^2\ln L}{\partial p^2} = p^{-2}s_{\bullet}+(1-p)^{-2}\sum_{i=1}^{m}(i-1)s_i+(N-s_{\bullet})\left(m(m-1)(1-p)^{-2}-m^2(1-p)^{-2}\right)$$

$$= p^{-2}s_{\bullet}+(1-p)^{-2}\left(\sum_{i=1}^{m}(i-1)s_i + m(N-s_{\bullet})\right)$$

$$= p^{-2}s_{\bullet}+(1-p)^{-2}\left(mN - \sum_{i=1}^{m}(m-i+1)s_i\right) .$$

These are the same formulas as Moran (1951, p. 309) reports.

4. Variance of Total Biomass Estimate

Notation: $\hat{\underline{N}} \equiv$ vector of estimated subpopulation totals.

$\underline{N} \equiv$ vector of subpopulation totals.

$\hat{\underline{w}} \equiv$ vector of estimated mean weights.

$\hat{w}_i \equiv$ estimated mean weight for subpopulation i .

$\underline{w} \equiv E(\hat{\underline{w}})$.

$V^* \equiv E\Big((\hat{N}-N)(\hat{N}-N)'\Big)$.

$D_w \equiv$ diagonal matrix with $\sigma^2_{w_i}$'s, the variances of the $\hat{w}_i$'s as the diagonal elements.

Formulas: $E(\hat{\underline{w}}'\hat{\underline{N}}|\hat{\underline{w}}) = \underline{w}'\underline{N}$,

$V(\hat{\underline{w}}'\underline{N}) = \underline{N}'D_{\underline{w}}\underline{N}$,

$V(\hat{\underline{w}}'\hat{\underline{N}}|\hat{\underline{w}}) = \hat{\underline{w}}'V^*\hat{\underline{w}}$,

$E(\hat{\underline{w}}'V^*\hat{\underline{w}}) = \underline{w}'V^*\underline{w} + tr(V^*D_w)$.

Using the above formulas, the variance of the estimated total biomass $\hat{\underline{w}}'\hat{\underline{N}}$ is:

$$\begin{aligned} Var(\hat{\underline{w}}'\hat{\underline{N}}) &= Var\Big(E(\hat{\underline{w}}'\hat{\underline{N}}|\hat{\underline{w}})\Big) + E\Big(Var(\hat{\underline{w}}'\hat{\underline{N}}|\hat{\underline{w}})\Big) \\ &= \underline{N}'D_{\underline{w}}\underline{N} + \underline{w}'V^*\underline{w} + tr(V^*D_w) \\ &= \sum_{i=1}^{k}\Big(\sigma^2_{w_i}(N_i^2 + V_{ii}) + \sum_{j=1}^{k} w_i w_j v_{ij}\Big) \end{aligned}$$

An estimated sampling variance of $(\hat{\underline{w}}'\hat{\underline{N}})$ is $Var(\hat{\underline{w}}'\hat{\underline{N}})$ with estimates replacing all parameters.

An example

The researcher working with Daphnia magna worked with $\theta = 0.20$, six size categories and three catches with a variety of different treatments on the populations. Table 1 is an example from his reported data.

Table 1

	s_{ij}	1	2	3	4	5	6	$s_{i.}$
		Size Category						
Catch	1	42	43	25	34	84	95	323
	2	29	27	9	24	59	69	217
	3	13	8	6	2	8	31	68
	$s_{.j}$	84	78	40	60	151	195	608

This data results in the following estimates using formulas (3), (4) and (5):

$\hat{p}$	$\hat{N}_.$	$\hat{N}_1$	$\hat{N}_2$	$\hat{N}_3$	$\hat{N}_4$	$\hat{N}_5$	$\hat{N}_6$
.491	3500	483.5	449.0	230.5	345.5	869.0	1122.5

For variance estimation the terms in V^{-1} become:

$$
\begin{aligned}
a &= 1847.9 \\
b &= 0.1881 \\
d_1 &= 0.00043 \\
d_2 &= 0.00047 \\
d_3 &= 0.00091 \\
d_4 &= 0.00061 \\
d_5 &= 0.00024 \\
d_6 &= 0.00019
\end{aligned}
$$

The resulting $\hat{V}^*$ matrix is

$$
\hat{V}^* = \begin{bmatrix}
2457 & 139 & 72 & 107 & 269 & 347 \\
139 & 2272 & 67 & 100 & 250 & 323 \\
72 & 67 & 1137 & 51 & 129 & 166 \\
107 & 100 & 51 & 1727 & 192 & 248 \\
269 & 250 & 129 & 192 & 4625 & 623 \\
347 & 323 & 166 & 248 & 623 & 6153
\end{bmatrix} .
$$

The average weights and their estimated standard deviations are recorded in Table 2.

Table 2

	Size Category					
	1	2	3	4	5	6
Weight (mg.)	0.1571	0.1351	0.0951	0.0564	0.0304	0.0159
s.d.	0.0081	0.0061	0.0046	0.0034	0.0061	0.0011

The resulting estimate of total biomass is

$$w'm = 222.29 \text{ mg.}$$

with estimated sampling variance of 203.6 . There was replication of treatments so variance of estimated treatment means can be estimated empirically.

Summary

The method of population size estimation by removal sampling is generalized allowing subsampling within catches at a known rate and partitioning of the population into subpopulations. This generalization is motivated by a desire to estimate the total biomass of a population and the sampling variance of that estimate. Previous work has involved approximations of the maximum likelihood estimates while it is possible to write exact solution equations. Both exact and approximate equations are presented.

Acknowledgement

I would like to acknowledge the help and guidance of D. S. Robson in work here presented.

References

Heimbuch, Doug (1977). M. S. Thesis, Cornell University.

Heimbuch, Doug (1977). Personal communication.

Moran, P. A. P. (1951). A mathematical theory of animal trapping. Biometrika 38, 307-311.

Searle, S. R. (1966). Matrix Algebra for the Biological Sciences, Wiley and Sons, New York.

Seber, G. A. F. (1973). The Estimation of Animal Abundance, Griffin, London.

Zippin, C. (1956). An evaluation of the removal method of estimating animal populations. Biometrics 12, 163-169.

Zippin, C. (1958). The removal method of population estimation. J. Wildlife Management 22, 82-90.